

REMARKS

Claims 1, 3-6, and 8-14 are pending in the present application. Claims 2 and 7 were canceled. Claims 1, 3, 4, 6, 11, and 12 were amended. Reconsideration of the claims is respectfully requested.

Amendments were made to the specification to correct errors and to clarify the specification. No new matter has been added by any of the amendments to the specification.

I. 35 U.S.C. § 102, Anticipation, Claims 1, 4-7, and 10-12

The Examiner has rejected claims 1, 4-7, and 10-12 under 35 U.S.C. § 102 as being anticipated by *Shriberg et al.* (ELIZABETH SHRIBERG ET AL., SPEECH COMMUNICATION 32 (2000) 127-154). This rejection is respectfully traversed.

A prior art reference anticipates the claimed invention under 35 U.S.C. § 102 only if every element of a claimed invention is identically shown in that single reference, arranged as they are in the claims. (*In re Bond*, 910 F.2d 831, 832, 15 U.S.P.Q.2d 1566, 1567 (Fed. Cir. 1990)). All limitations of the claimed invention must be considered when determining patentability. (*In re Lowry*, 32 F.3d 1579, 1582, 32 U.S.P.Q.2d 1031, 1034 (Fed. Cir. 1994)). Anticipation focuses on whether a claim reads on the product or process a prior art reference discloses, not on what the reference broadly teaches. (*Kalman v. Kimherly-Clark Corp.*, 713 F.2d 760, 218 U.S.P.Q. 781 (Fed. Cir. 1983)).

A. Amended independent claim 1 of the present invention, which is representative of amended independent claims 11 and 12, reads as follows:

1. A method for the segmentation of an audio stream into semantic or syntactic units wherein the audio stream is provided in a digitized format; comprising the steps of:

determining a fundamental frequency for the digitized audio stream;

detecting changes of the fundamental frequency in the audio stream, wherein detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value;

determining candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency;

extracting and combining a plurality of prosodic features in the neighborhood of the candidate boundaries; and

determining boundaries for the semantic or syntactic units depending on the combined plurality of prosodic features.

With regard to claim 1, the Examiner stated:

Regarding claims 1 and 11-12, Shriberg et al. discloses a method, a computer usable medium having computer readable program code, and a digital audio processing system for the segmentation of an audio stream into semantic or syntactic units wherein the audio stream is provided in a digitized format, comprising the steps of:

determining a fundamental frequency for the digitized audio stream (*Section 2.1.2.3 on page 133*);

detecting changes of the fundamental frequency in the audio stream (*pages 134-135, refer to figure 4*);

determining candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency (*pages 134-135*);

extracting at least one prosodic feature in the neighborhood of the candidate boundaries (*pages 130-131*);

determining boundaries for the semantic or syntactic units depending on the at least one prosodic feature (*pages 134-135, F0 is a prosodic feature*).

(*Office Action*, dated September 21, 2004, Page 2 and 3)

Even though *Shriberg* teaches prosody-based automatic segmentation of speech into sentences and topics (*Shriberg*, Title), *Shriberg* does not identically teach every element of Applicants' present invention as recited in amended independent claim 1. Examiner Vo cites in the *Office Action*, *Shriberg*, Page 131, Section 2.1.2, with regard to "...the extracting step involves extracting at least two prosodic features and combining the at least two prosodic features...." (*Office Action*, Page 3). That particular section cited by the Examiner teaches that the prosodic "[f]eatures are **grouped into broad feature classes** based on the kinds of measurements involved, and the type of prosodic behavior they were designed to capture." (Emphasis added) (*Shriberg*, Page 131, Section 2.1.2). In other words, *Shriberg* teaches that the prosodic features are merely placed in

broad classes according to their measurements or behavior. *Shriberg* does not teach, nor does it mention the desirability of combining extracted prosodic features to determine semantic and syntactic units in the segmentation process.

In contrast, claim 1 of the current invention recites extracting and combining a plurality of prosodic features in the neighborhood of the candidate boundary for determining the semantic and syntactic units. In other words, Applicants' invention combines at least two extracted prosodic features from the audio stream's fundamental frequency boundaries for determining the semantic and syntactic units in the segmentation process. *Shriberg* does not teach the extraction and combination of prosodic features in the segmentation process as recited in claim 1 of the present invention. Therefore, *Shriberg* does not teach this recited element in claim 1.

Furthermore, Examiner Vo stated in the Office Action that, "*Shriberg et al.* does not disclose a method for providing a threshold value for the voicedness of the fundamental frequency estimates." (*Office Action*, Page 4). Applicants agree with Examiner Vo that *Shriberg* does not teach this feature as recited in amended claim 1 of the current invention.

Moreover, the *Shriberg* method employs the use of a speech recognizer in the segmentation process, whereas the Applicants' present invention does not. *Shriberg* states that:

... [W]e created parallel prosodic databases for both corpora, and used the same machine learning approach for prosodic modeling in all cases. We look at results for both true words, and words as hypothesized by a speech recognizer. Both conditions provide informative data points. *Using recognized words allows comparison of degradation of the prosodic model to that of a language model, and also allows us to assess realistic performance of the prosodic model when word boundary information must be extracted based on incorrect hypotheses rather than forced alignments.* (Emphasis added).

(*Shriberg*, Page 30, Section 1.4).

As the above passage clearly indicates, a speech recognizer is used in *Shriberg's* audio stream segmentation process for the purpose of comparison with the prosodic model. Conversely, the present invention teaches away from the use of a speech

recognizer. For example, Applicants' current invention teaches that there is no need for a speech recognizer to be invoked in the segmentation process. (*Application*, Page 22, lines 19 and 20). Thus, Applicants' current invention does not recite nor teach the use of a speech recognizer in the segmentation process as does the method in *Shriberg*.

Finally, *Shriberg* teaches that "[a] crucial step in processing speech audio data for information extraction, topic detection, or browsing/playback is to segment the input into sentence and topic units." (Emphasis added) (*Shriberg*, Abstract). The immediately preceding passage is indicative of *Shriberg*'s teaching with regard to speech audio data. The speech audio data in *Shriberg* is in an analog format from two speech corpora, Broadcast News and Switchboard. (*Id.*). *Shriberg* does not teach, nor mention the desirability of providing the audio stream in a digitized format. In contrast, claim 1 of the current invention recites that the audio stream is provided in a digitized format. Therefore, *Shriberg* does not teach this recited claim 1 limitation either.

As a result of the foregoing arguments, *Shriberg* does not identically teach each and every element recited in the present invention. Accordingly, Applicants respectfully urge that the rejection of amended independent claims 1, 11, and 12 under 35 U.S.C. § 102 be withdrawn.

B. In view of the arguments contained in Section A above, Applicants respectfully submit that each element of amended independent claims 1, 11, and 12 are not identically taught by *Shriberg*. Claims 3-6, 8-10, 13, and 14 are dependent claims depending on the aforementioned independent claims. Applicants have already demonstrated claims 1, 11, and 12 to be in condition for allowance. Therefore, Applicants respectfully submit that claims 3-6, 8-10, 13, and 14 are also allowable, at least by virtue of their dependence on allowable claims.

Furthermore, dependent claims 3, 5, 8, 9, 13 and 14 contain features not taught by the *Shriberg* reference.

C. Dependent method claims 3, which is representative of dependent digital audio processing system claim 13, reads as follows:

3. The method according to claim 1, wherein defining an index function for the fundamental frequency having a value = 0 if the voicedness of the fundamental frequency is lower than the threshold value and having a value = 1 if the voicedness of the fundamental frequency is higher than the threshold value.

With regard to claims 3 and 13, the Examiner stated:

Regarding claims 2-3 and 13, **Shriberg et al. does not disclose a method for providing a threshold value for the voicedness of the fundamental frequency estimates and determining whether the voicedness of fundamental frequency estimates is lower than the threshold value, and for defining an index function for the fundamental frequency having a value = 0 if the voicedness of the fundamental frequency is lower than the threshold value and having a value = 1 if the voicedness of the fundamental frequency is higher than the threshold value.** (Emphasis added).

(Office Action, Page 4).

Applicants agree with Examiner Vo that *Shriberg* does not teach this feature as recited in dependent claims 3 and 13 of the current invention.

D. Dependent claim 5 of the present invention reads as follows:

5. The method according to claim 4, wherein the environment is a time period between 500 and 4000 milliseconds.

With regard to claim 5, the Examiner stated:

Regarding claims 4-7 and 10, *Shriberg et al.* further disclose a method for extracting at least one prosodic feature in an environment of the audio stream where the value of the index function is equal 0 (*section 2.1.1 on page 130 discusses feature extraction of both voice and unvoiced portions*), **that the environment is a time period between 500 and 4000 milliseconds (Section 2.1.1 on page 130)**, at least one prosodic feature is represented by the fundamental frequency (*Section 2.1.1, page 130*), the extracting step involves extracting at least two prosodic features and combining the at least two prosodic features (*Section 2.1.2, page 131*), and

a step of performing a prosodic feature classification based on a predetermined classification tree (*section 2.1.2 on page 131, grouping features*). (Emphasis added).

(Office Action, Page 3).

Shriberg teaches that "...for each inter-word boundary, we looked at prosodic features of the word immediately preceding and following the boundary, or alternatively within a window of 20 frames (200 ms, a value empirically optimized for this work) before and after the boundary." (Emphasis added) (*Shriberg*, Page 130, Section 2.1.1.). In other words, *Shriberg* is very specific as to the time period employed in the audio stream segmentation process. The only value empirically optimized for *Shriberg's* method is 200 ms.

However, Applicants' present invention recites in claim 5 that the environment for extracting the two prosodic features from the candidate boundaries is a time period between 500 and 4000 milliseconds (ms). Thus, not only does *Shriberg* teach just one time value (200 ms) in the segmentation process, but that value does not even fall within the 500 to 4000 ms range recited in claim 5 of the current invention. Hence, *Shriberg* does not teach the feature recited in dependent claim 5 of the Applicants' invention.

Accordingly, Applicants respectfully urge that the rejection of dependent claim 5 under 35 U.S.C. § 102 be withdrawn.

E. Dependent method claims 8 and 9, which are representative of dependent digital audio processing system claim 14, reads as follows:

8. The method according to claim 1, further comprising first detecting speech and non-speech segments in the digitized audio stream and performing the steps of claim 1 thereafter only for detected speech segments.

9. The method according to claim 8, wherein the detecting of speech and non-speech segments comprises utilizing the signal energy or signal energy changes, respectively, in the audio stream.

With regard to claims 8, 9, and 14, the Examiner stated:

Regarding claim 8, *Shriberg et al.* does not disclose a method that first detects speech and non-speech segments in the digitized audio stream

and performs the steps of claim 1 thereafter only for detected speech segments.

Regarding claims 9 and 14, the modified Shriberg et al., as applied to claims 8 and 13 above, fails to disclose a method of detecting of speech and non-speech segments comprises utilizing the signal energy or signal energy changes, respectively, in the audio stream.

(Office Action, Pages 5 and 6).

Applicants agree with Examiner Vo that *Shriberg* does not teach the recited features in dependent claims 8, 9, and 14 of the present invention. Accordingly, Applicants respectfully submit once again that *Shriberg* does not identically teach each and every element recited in the current invention. Therefore, Applicants respectfully urge that the rejection of the present invention under 35 U.S.C. § 102 be withdrawn.

II. 35 U.S.C. § 103, Obviousness, Dependent Claims 2-3, 8, and 13

The Examiner has rejected dependent claims 2-3, 8, and 13 under 35 U.S.C. § 103 as being unpatentable over *Shriberg et al.* (ELIZABETH SHRIBERG ET AL., SPEECH COMMUNICATION 32 (2000) 127-154) in view of *Yeldener et al.* (U.S. Patent No. 5,774,837). This rejection is respectfully traversed.

The Examiner bears the burden of establishing a *prima facie* case of obviousness based on the prior art when rejecting claims under 35 U.S.C. § 103. (*In re Fritch*, 972 F.2d 1260, 23 U.S.P.Q.2d 1780 (Fed. Cir. 1992)). For an invention to be *prima facie* obvious, the prior art must teach or suggest all claim limitations. (*In re Royka*, 490 F.2d 981, 180 USPQ 580 (CCPA 1974)).

Canceled dependent claim 2 had its claim language incorporated into independent claims 1, 11, and 12 of the present invention. Amended independent claim 1 of the present invention, which is representative of amended independent claims 11 and 12, reads as follows:

1. A method for the segmentation of an audio stream into semantic or syntactic units wherein the audio stream is provided in a digitized format, comprising the steps of:

determining a fundamental frequency for the digitized audio stream;

detecting changes of the fundamental frequency in the audio stream, wherein detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value;

determining candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency;

extracting and combining a plurality of prosodic features in the neighborhood of the candidate boundaries; and

determining boundaries for the semantic or syntactic units depending on the combined plurality of prosodic features.

With regard to claims 2, 3, and 13, the Examiner stated:

Regarding claims 2-3 and 13, Shriberg et al. does not disclose a method for providing a threshold value for the voicedness of the fundamental frequency estimates and determining whether the voicedness of fundamental frequency estimates is lower than the threshold value, and for defining an index function for the fundamental frequency having a value = 0 if the voicedness of the fundamental frequency is lower than the threshold value and having a value = 1 if the voicedness of the fundamental frequency is higher than the threshold value.

However, Yeldener et al. teaches a method for providing a threshold value for the voicedness of the fundamental frequency estimates and determining whether the voicedness of fundamental frequency estimates is lower than the threshold value (*col. 15, ln. 1 to col. 16, l. 14*), and for defining an index function for the fundamental frequency is lower than the threshold value and having a value = 1 if the voicedness of the fundamental frequency is higher than the threshold value (*col. 14, ln. 4-55, the goal is to use 0 and 1 to represent for unvoiced and voiced portions, respectively*).

Since Shriberg et al. and Yeldener et al. are analogous art because they are from the same filed of endeavors, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Shriberg et al. by incorporating the teaching of Yeldener et al. in order to enable the system to pay more coding emphasis on the voice portion than unvoiced portion to reduce processing time and increase transmission rate.

(*Office Action*, Pages 4 and 5, Section 7).

Shriberg does teach prosody-based automatic segmentation of speech into sentences and topics. (*Shriberg*, Tile). However, *Shriberg* does not teach or suggest all claim limitations of Applicants' present invention as recited in amended independent claim 1. As argued in Section A above, *Shriberg* does not teach or suggest combining extracted prosodic features to determine semantic and syntactic units in the segmentation process of a digitized audio stream. In contrast, claim 1 of the current invention recites extracting and combining a plurality of prosodic features in the neighborhood of the candidate boundary for determining the semantic and syntactic units. Thus, *Shriberg* does not teach this recited feature in claim 1.

In addition, all of the Section I arguments are applicable herein to further demonstrate that *Shriberg* does not teach or suggest Applicants' present invention recited in the current claims. Consequently, *Shriberg* does not teach or suggest all claim limitations of the current invention.

But, Examiner Vo rejected claims 2, 3, and 13 under 35 U.S.C. § 103 as being unpatentable over *Shriberg* in view of *Yeldener*. *Yeldener* is a method for providing encoding and decoding of speech signals using voicing probability determination. (*Yeldener*, Abstract). Even though *Yeldener* teaches "a method for providing a threshold value for the voicedness of the fundamental frequency estimates and determining whether the voicedness of fundamental frequency estimates is lower than the threshold value (col. 15, ln. 1 to col. 16, l. 14), and for defining an index function for the fundamental frequency is lower than the threshold value and having a value = 1 if the voicedness of the fundamental frequency is higher than the threshold value (col. 14, ln. 4-55, the goal is to use 0 and 1 to represent for unvoiced and voiced portions, respectively)" (*Office Action*, Page 4), *Yeldener* does not teach or suggest combining extracted prosodic features to determine semantic and syntactic units in the segmentation process of a digitized audio stream as recited in amended claim 1 of Applicants' current invention.

Yeldener teaches that:

...the input speech signal is represented as a sequence of time segments of predetermined length. For each input segment a determination is made as to detect the presence and estimate the frequency of the pitch F_0 of the speech signal within the time segment. Next, on the basis of the estimated

pitch is determined the probability that the speech signal within the segment contains voiced speech patterns.

(*Yeldener*, Column 4, lines 25-32).

As the passage above clearly indicates, *Yeldener* teaches that only the prosodic feature of pitch is utilized to determine the segments that contain voiced speech patterns. In other words, only one prosodic feature is employed in the method of *Yeldener*. As a result, *Yeldener* cannot combine a plurality of extracted prosodic features to determine semantic and syntactic units in the segmentation process of a digitized audio stream as recited in amended claim 1 of Applicants' current invention.

Hence, since neither *Shriberg* nor *Yeldener* teach the recited claim limitation of extracting and combining a plurality of prosodic features, then the combination of *Shriberg* and *Yeldener* cannot teach or suggest all the claim limitations as recited in amended independent claim 1. Accordingly, Applicants respectfully urge that the rejection of claims under 35 U.S.C. § 103 as being unpatentable over *Shriberg* in view of *Yeldener* be withdrawn.

III. 35 U.S.C. § 103, Obviousness, Dependent Claims 9 and 14

The Examiner has rejected dependent claims 9 and 14 under 35 U.S.C. § 103 as being unpatentable over *Shriberg et al.* (ELIZABETH SHRIBERG ET AL., SPEECH COMMUNICATION 32 (2000) 127-154) in view of *Yeldener et al.* (U.S. Patent No. 5,774,837), as applied to claims 8 and 13 above, and further in view of *Eryilmaz* (U.S. Patent No. 5,867,574). This rejection is respectfully traversed.

With regard to *Shriberg*, the arguments in Sections I and II above illustrate that *Shriberg* does not teach or suggest combining extracted prosodic features to determine semantic and syntactic units in the segmentation process of a digitized audio stream as recited in the amended independent claims of Applicants' present invention. Therefore, *Shriberg* does not teach or suggest all claim limitations of the current invention.

With regard to *Yeldener*, the arguments in Section II above reveal that *Yeldener* does not teach or suggest combining extracted prosodic features to determine semantic and syntactic units in the segmentation process of a digitized audio stream as recited in

the amended independent claims of Applicants' present invention. Thus, *Yeldener* does not teach or suggest all claim limitations of the current invention.

With regard to *Eryilmaz*, *Eryilmaz* is "an improved voice detection method used in the half-duplex speakerphone that operates in a transmit mode, a receive mode, and a silence mode. In general, the improved voice detection method includes measuring the voice energy level for each frame of sampled data in a speech signal." (*Eryilmaz*, Column 3, lines 51-56). In other words, *Eryilmaz* merely measures voice energy levels in a speakerphone's speech signal. There is no reference of any type in *Eryilmaz* with regard to prosodic features being used to analyze the speech signal. Consequently, *Eryilmaz* cannot teach or suggest the combining of a plurality of extracted prosodic features to determine semantic and syntactic units in the segmentation process of a digitized audio stream as recited in amended claim 1 of Applicants' current invention. Thus, *Eryilmaz* does not teach or suggest all claim limitations of the current invention.

As a result, since none of the above cited references teach the recited claim limitation of extracting and combining a plurality of prosodic features, then the combination of *Shriberg*, *Yeldener*, and *Eryilmaz* cannot teach or suggest all the claim limitations as recited in the amended independent claims of the present invention. Accordingly, Applicants respectfully urge that the rejection of claims under 35 U.S.C. § 103 as being unpatentable over *Shriberg*, in view of *Yeldener*, in further view of *Eryilmaz* be withdrawn.

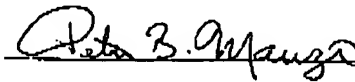
IV. Conclusion

It is respectfully urged that the subject application is patentable over the cited references and is now in condition for allowance.

The Examiner is invited to call the undersigned at the below-listed telephone number if in the opinion of the examiner such a telephone conference would expedite or aid the prosecution and examination of this application.

DATE: 12-21-04

Respectfully submitted,



Peter B. Manzo
Reg. No. 54,700
Yee & Associates, P.C.
P.O. Box 802333
Dallas, TX 75380
(972) 385-8777
Attorney for Applicants